RESTAURANT LOCATION RESEARCH IN DUBLIN

IBM DATA SCIENCE CAPSTONE PROJECT PRESENTATION

JOSE HERNANDEZ

Dublin
June 2020

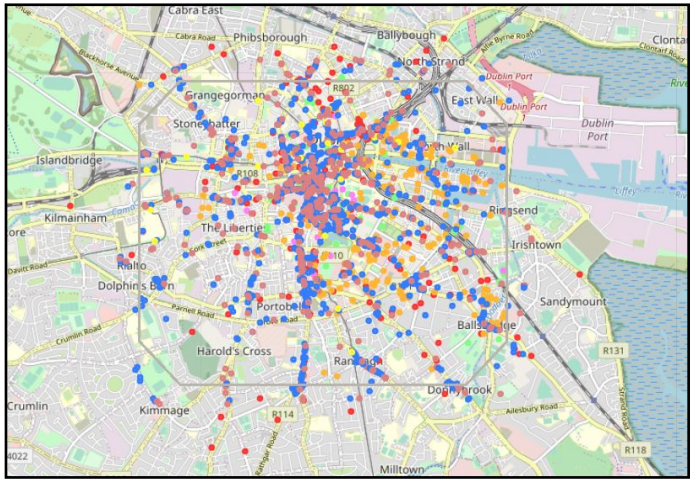
jh.hdez@gmail.com

INTRODUCTION

D M R D C INTRODUCTION

Dublin City

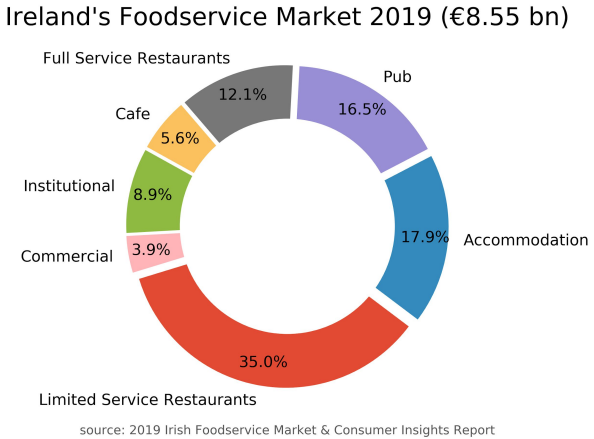
- capital of Ireland
- over 318 square kilometres
- 1 million people living in the city
- one of the highest ratios of tourists to locals in the world

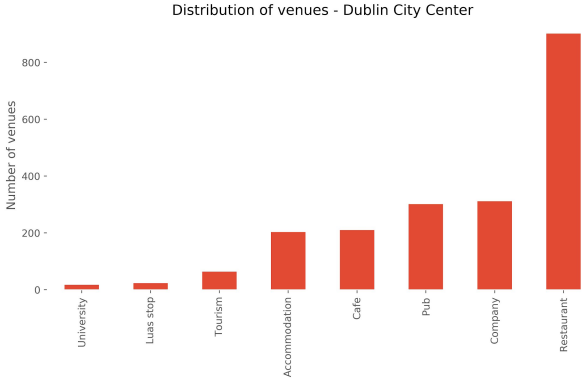


Tourism, Restaurant, Accommodation, University, Cafe, Pub, Companies, Luas stop, Project boundary

INTRODUCTION

INTRODUCTION





- Foodservice or ‘Out of Home’ is the term used to describe all food consumed and prepared out of home. Almost half of the market (47.1%) in Dublin is shared by Full Service Restaurants, and Limited Service Restaurants, like fast food businesses.
- There are over 800 restaurants in Dublin City, as shown in the chart

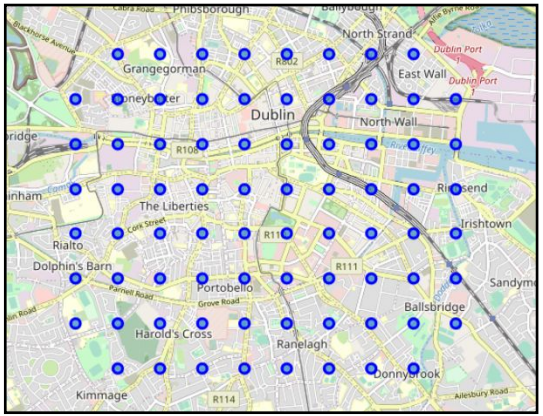
INTRODUCTION

Getting into the scope of this study, we can consider some questions, like if there is any relationship between restaurant success and the closeness to:

- ❏ the river Liffey that crosses the city
- ❏ a Luas tram station helping accessibility to your business
- ❏ accommodations, such as Bed & Breakfast, Hostels or Hotels
- ❏ tourist attractions and museums
- ❏ cafes
- ❏ pubs
- ❏ companies
- ❏ universities
- ❏ other restaurants

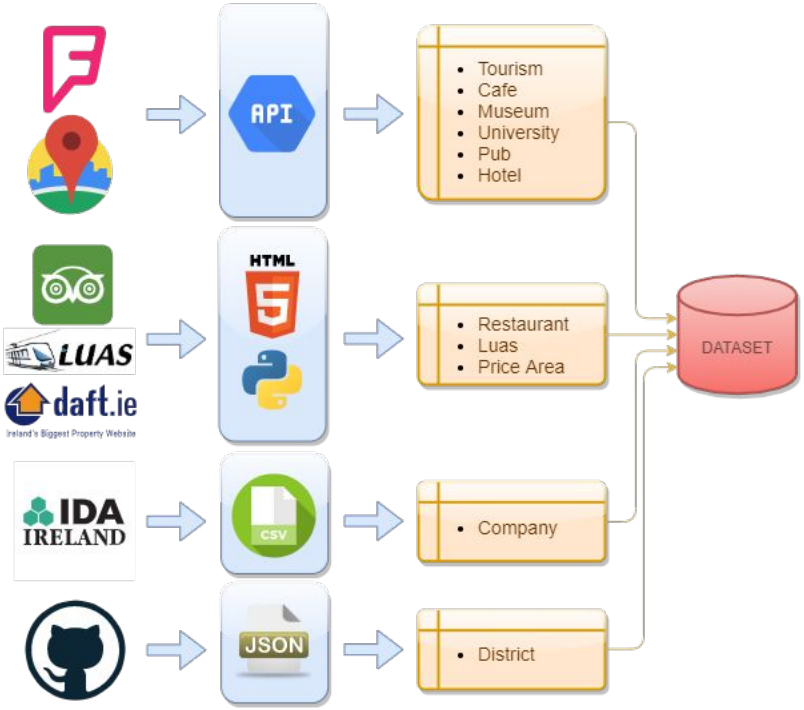
DATA

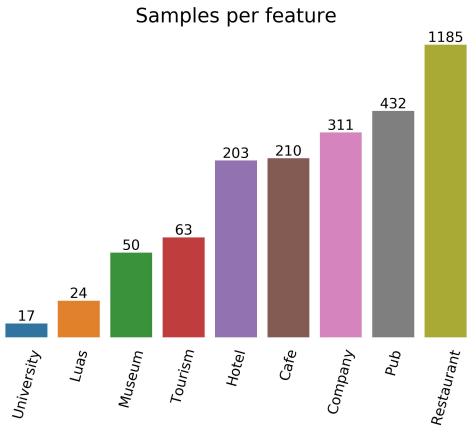
I D M R D C DATA



mesh of points on top of Dublin city due to asymmetrical boroughs shape, so we can use each point in order to explore the area leveraging the Foursquare API

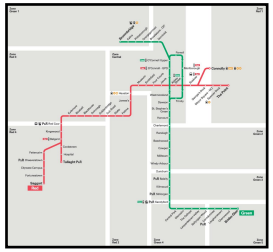
1 DATA ADQUISITION

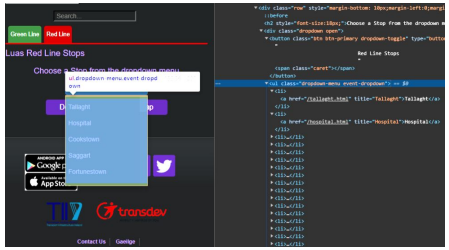




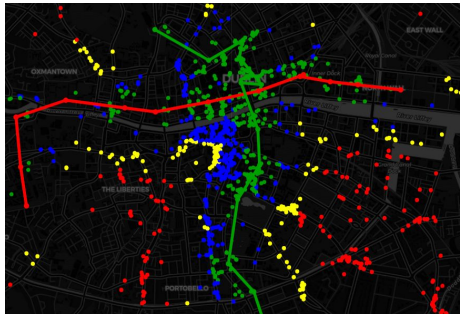
2 DATA ADQUISITION

IDMRDC DATA





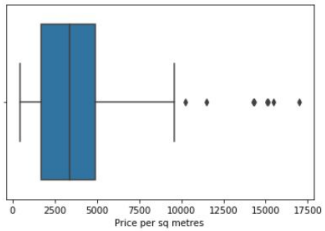
➢ Luas stops data has been collected by web scraping the official website

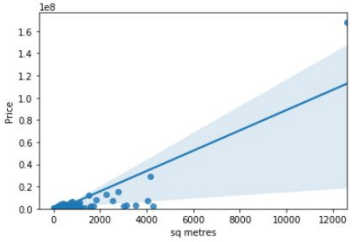


➢ Once data had been collected, we calculated the distance between each restaurant to the closest Luas stop, fixing the minimum distance as 100 meter

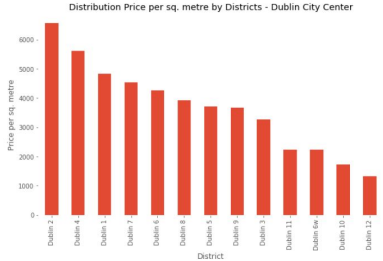
DATA CLEANING

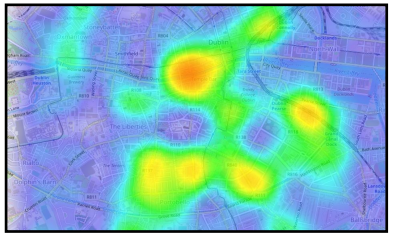
I D M R D C DATA





- identifying some outliers and remove them

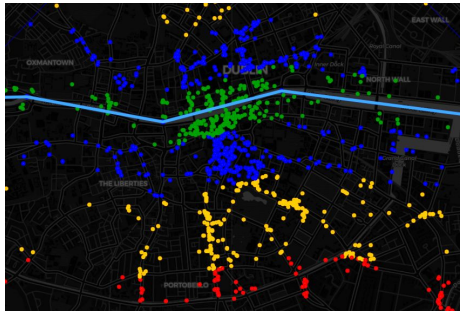




- Chart with the average price for each district and visualization on a heatmap

DATA GENERATION

Generation of dots on the river



➢ Restaurants proximity to the Liffey river

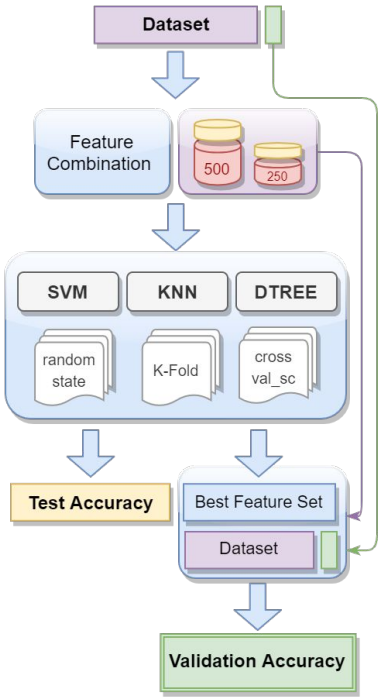
IDA Ireland

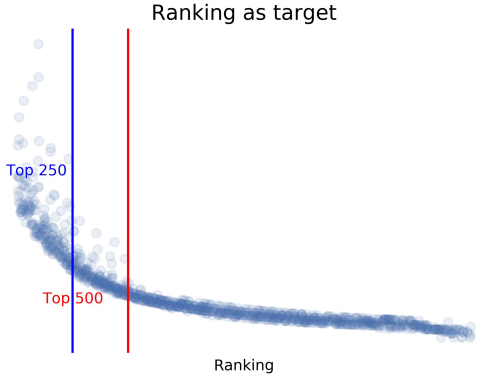
311 companies in the Tech Hub Ireland in Dublin City

➢ Clustering using K-means

METHODOLOGY

METHODOLOGY

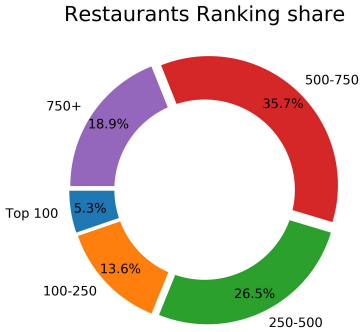


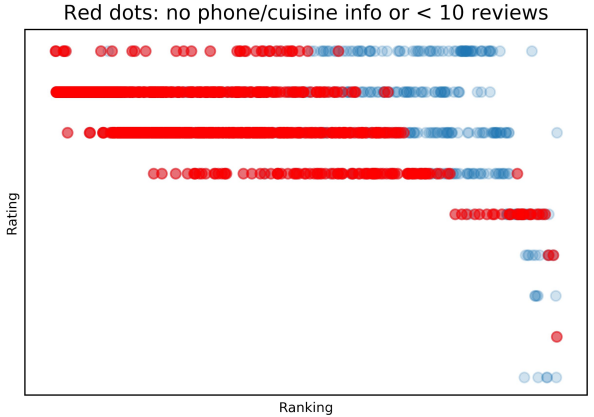


- The idea is to have a predictive model where we can foretell if a new location for a restaurant will share similarities with current Top 250 & Top 500 ranges
- Data Science is able to estimate a probability of business success related to the location
- This variable combined with other factors will provide an edge for a potential investment

1 EXPLORATORY DATA ANALYSIS

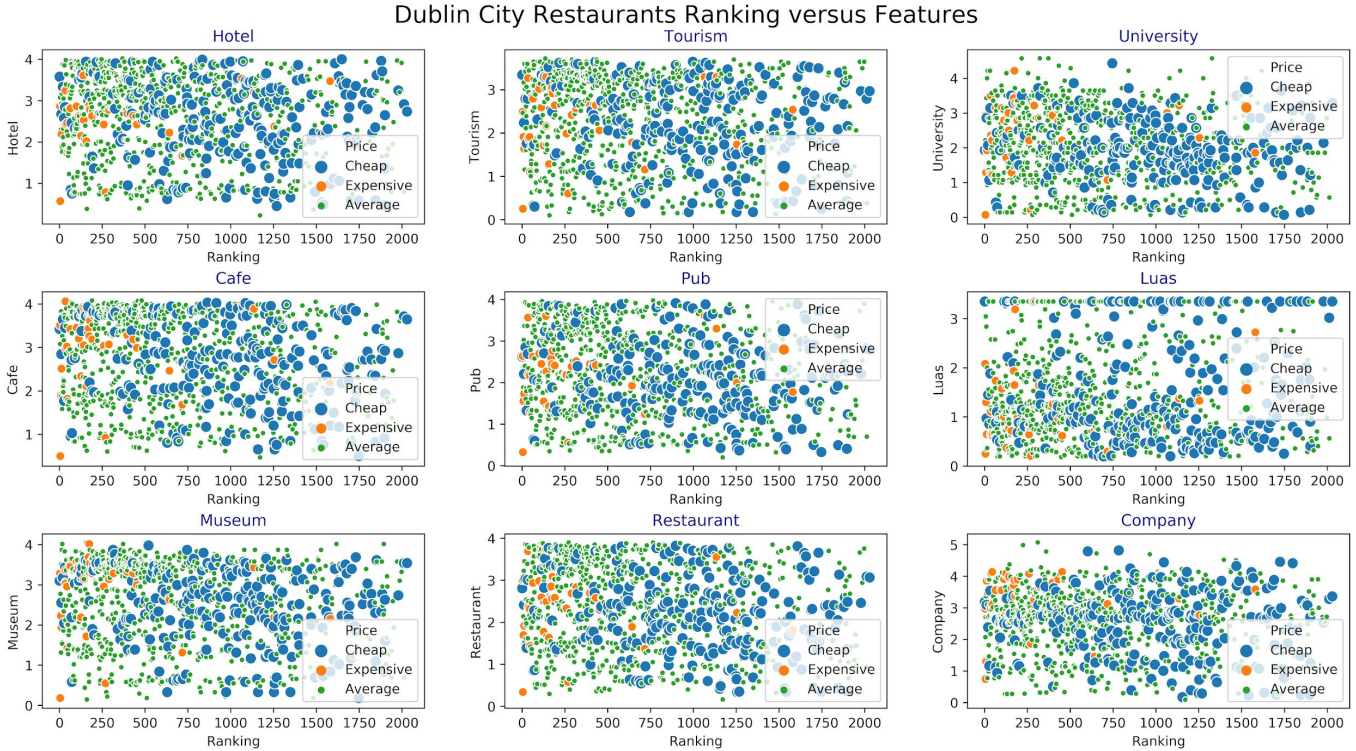
I D M R D C METHODOLOGY





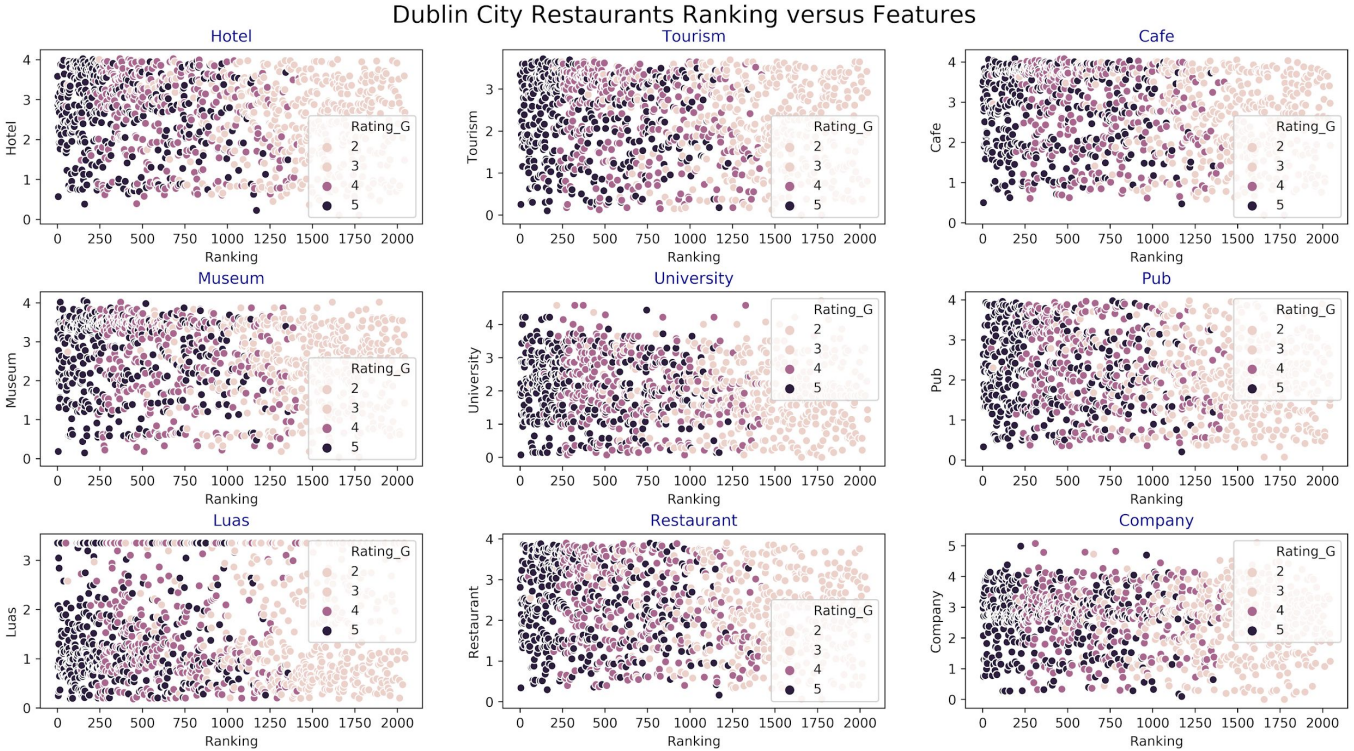
- After cleaning the data, we obtained the next rating and ranking distributions in the dataset to work with
- In order to filter some restaurant samples we need to remove unnecessary data
- As shown in the chart, red dots would be restaurants with no phone or cuisine information or less than 10 reviews

EXPLORATORY DATA ANALYSIS



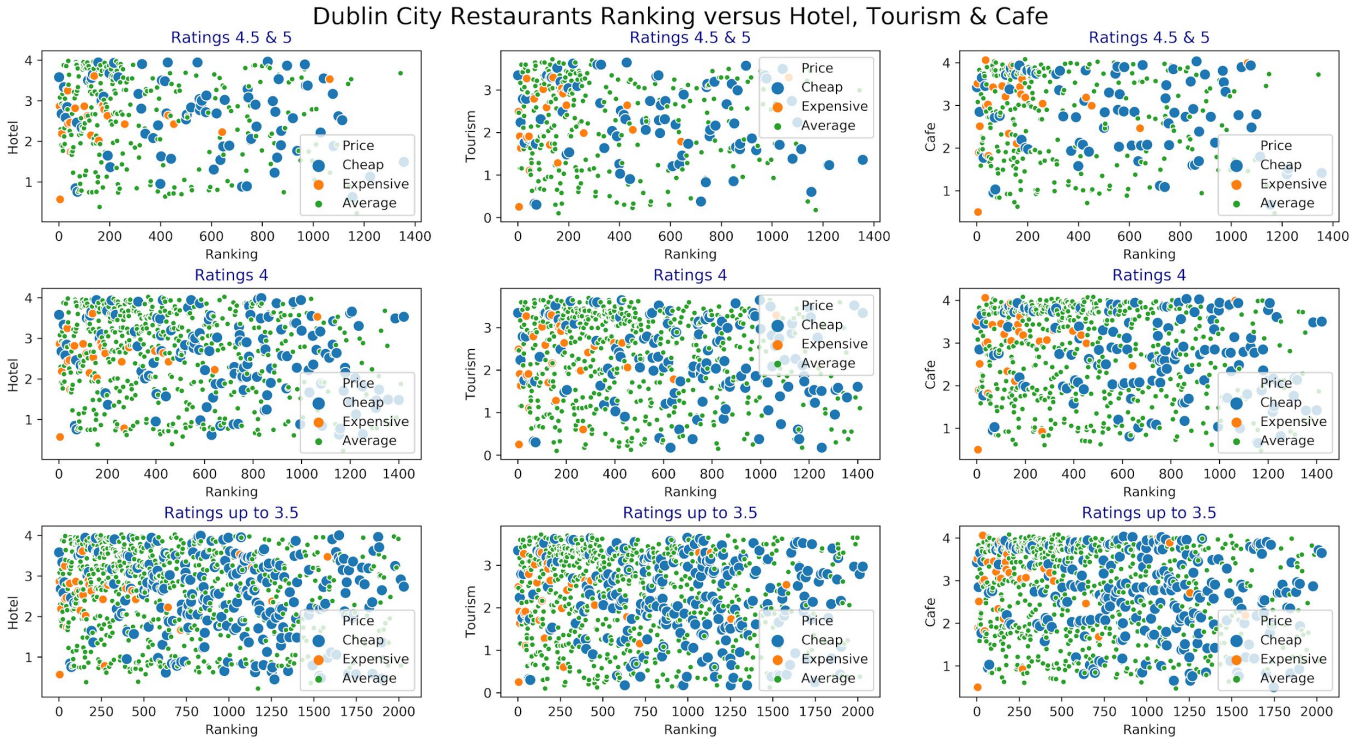
3 EXPLORATORY DATA ANALYSIS

IDMRDC METHODOLOGY



4 EXPLORATORY DATA ANALYSIS

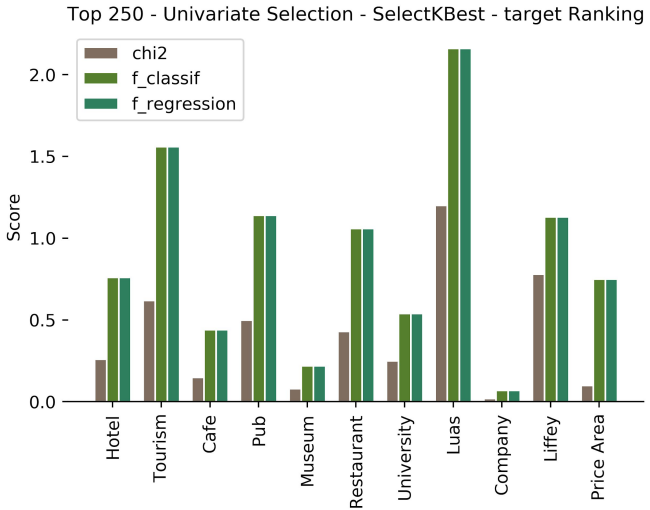
IDMRDC METHODOLOGY

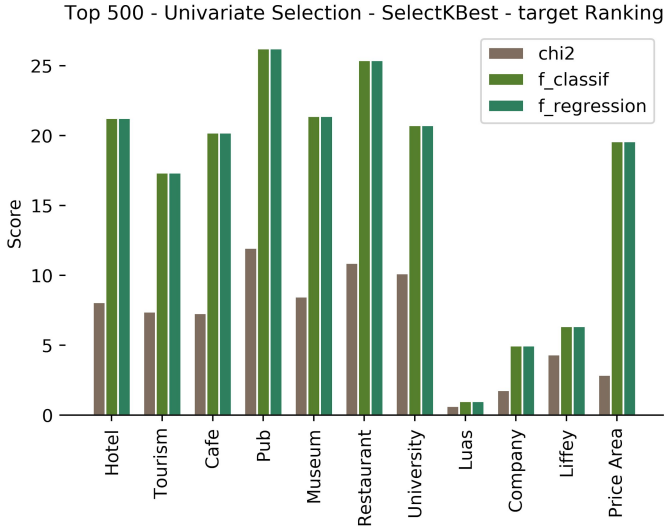


INFERENTIAL STATISTICS

I D M R D C METHODOLOGY

- UNIVARIATE STATISTICS
- FEATURE IMPORTANCE
- CORRELATION MATRIX
- SCIKIT ALGORITHMS



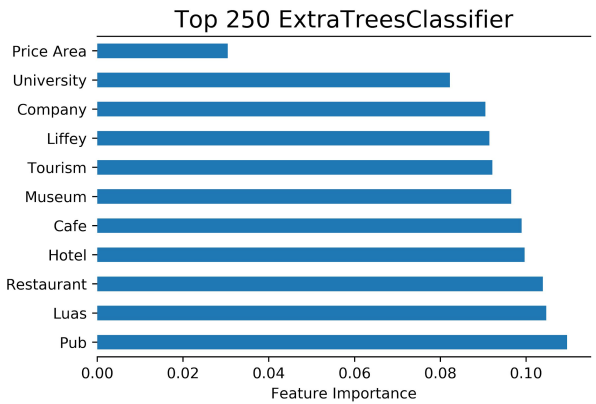


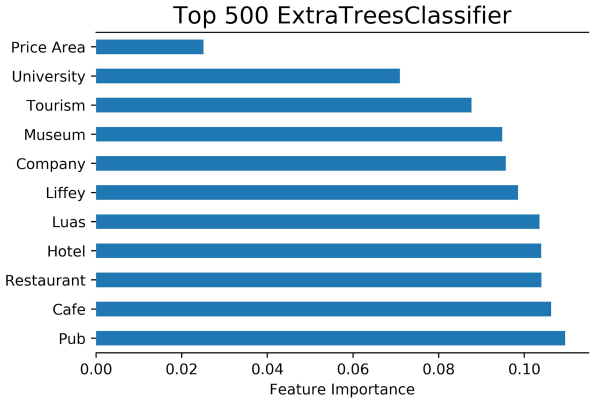
- chi2: Chi-squared stats of non-negative features
- f_classif: ANOVA F-value between label/feature
- f_regression: F-value between label/feature

INFERENTIAL STATISTICS

I D M R D C METHODOLOGY

- UNIVARIATE STATISTICS
- FEATURE IMPORTANCE
- CORRELATION MATRIX
- SCIKIT ALGORITHMS



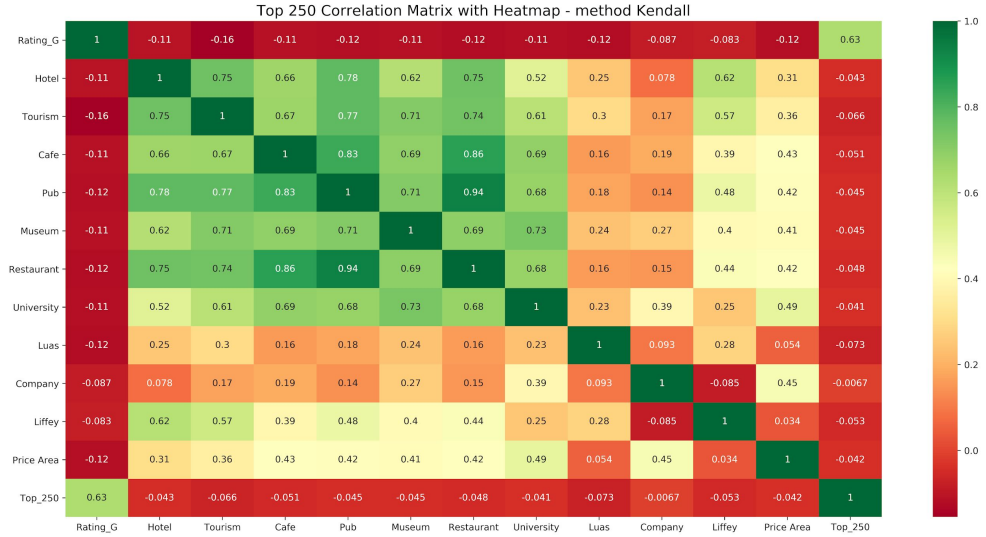


By averaging the estimates of predictive ability over several randomized trees one can reduce the variance of such an estimate and use it for feature selection.
(scikit-learn website)

INFERENTIAL STATISTICS

I D M R D C METHODOLOGY

- UNIVARIATE STATISTICS
- FEATURE IMPORTANCE
- CORRELATION MATRIX
- SCIKIT ALGORITHMS



For uncorrelated features, the optimal feature size is N-1 (N sample size)
As feature correlation increases, the optimal feature size is proportional to $\sqrt{(N)}$ for highly correlated features
However, a rule of thumb for a dataset of N samples would be N/10 features

INFERENTIAL STATISTICS

I D M R D C METHODOLOGY

UNIVARIATE STATISTICS

FEATURE IMPORTANCE

CORRELATION MATRIX

SCIKIT ALGORITHMS

Top_250

	Feature	Pearson	RFE	Logistics	Random Forest	LightGBM	Total
1	Restaurant	True	True	True	True	False	4
2	Luas	True	True	False	True	True	4
3	Tourism	True	True	True	False	False	3
4	Pub	True	False	False	True	True	3
5	Liffey	True	False	False	True	True	3
6	Hotel	True	True	False	True	False	3
7	Cafe	False	True	True	True	False	3
8	Museum	False	True	True	False	False	2
9	University	False	False	False	False	True	1
10	Company	False	False	False	False	True	1
11	Price Area	False	False	False	False	False	0

Top_500

	Feature	Pearson	RFE	Logistics	Random Forest	LightGBM	Total
1	Pub	True	True	True	True	True	5
2	Hotel	True	True	True	True	True	5
3	Cafe	True	True	True	True	True	5
4	Restaurant	True	True	True	True	False	4
5	Museum	True	True	True	False	False	3
6	Luas	False	False	False	True	True	2
7	Liffey	False	False	False	True	True	2
8	University	True	False	False	False	False	1
9	Tourism	False	True	False	False	False	1
10	Company	False	False	False	False	True	1
11	Price Area	False	False	False	False	False	0

- Pearson’s correlation coefficient (linear)
- Recursive Feature Elimination (REF)
- Lasso: SelectFromModel - Logistic Regression
- Tree-based: SelectFromModel - Random Forest Classifier
- LightGBM Classifier

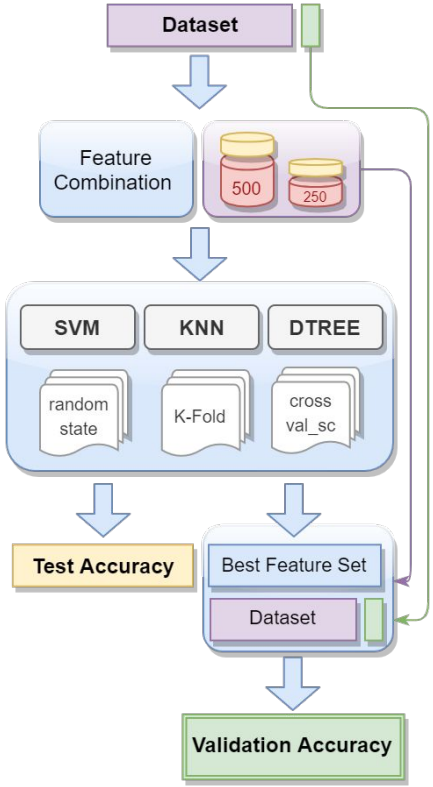
PREDICTION MODELS

I D M R D C METHODOLOGY

SUPPORT VECTOR MACHINE

K-NEAREST NEIGHBORS

DECISION TREE



- added two columns to the dataset filtering by ranking:
  - <= 500 for Top_250
  - <= 1500 for Top_500
- Out of the 11 features we generated 2047 different combinations of feature sets

Cross Validation Methods:

- Modifying the random_state parameter while we split the data into train and test sets
- K-Fold method with 10 folds
- Cross_val_score from scikit with 10 folds

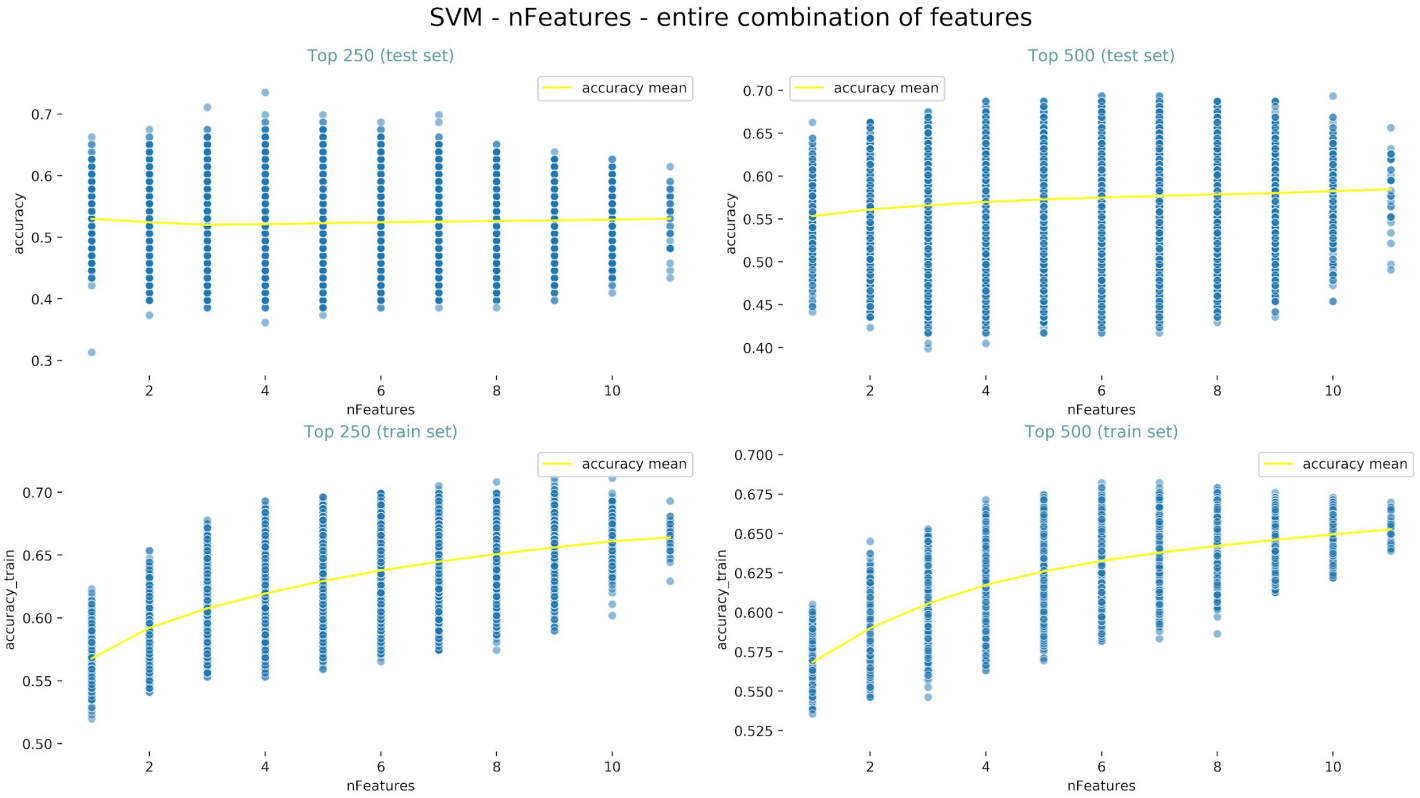
1 PREDICTION MODELS

I D M R D C METHODOLOGY

SUPPORT VECTOR MACHINE

K-NEAREST NEIGHBORS

DECISION TREE



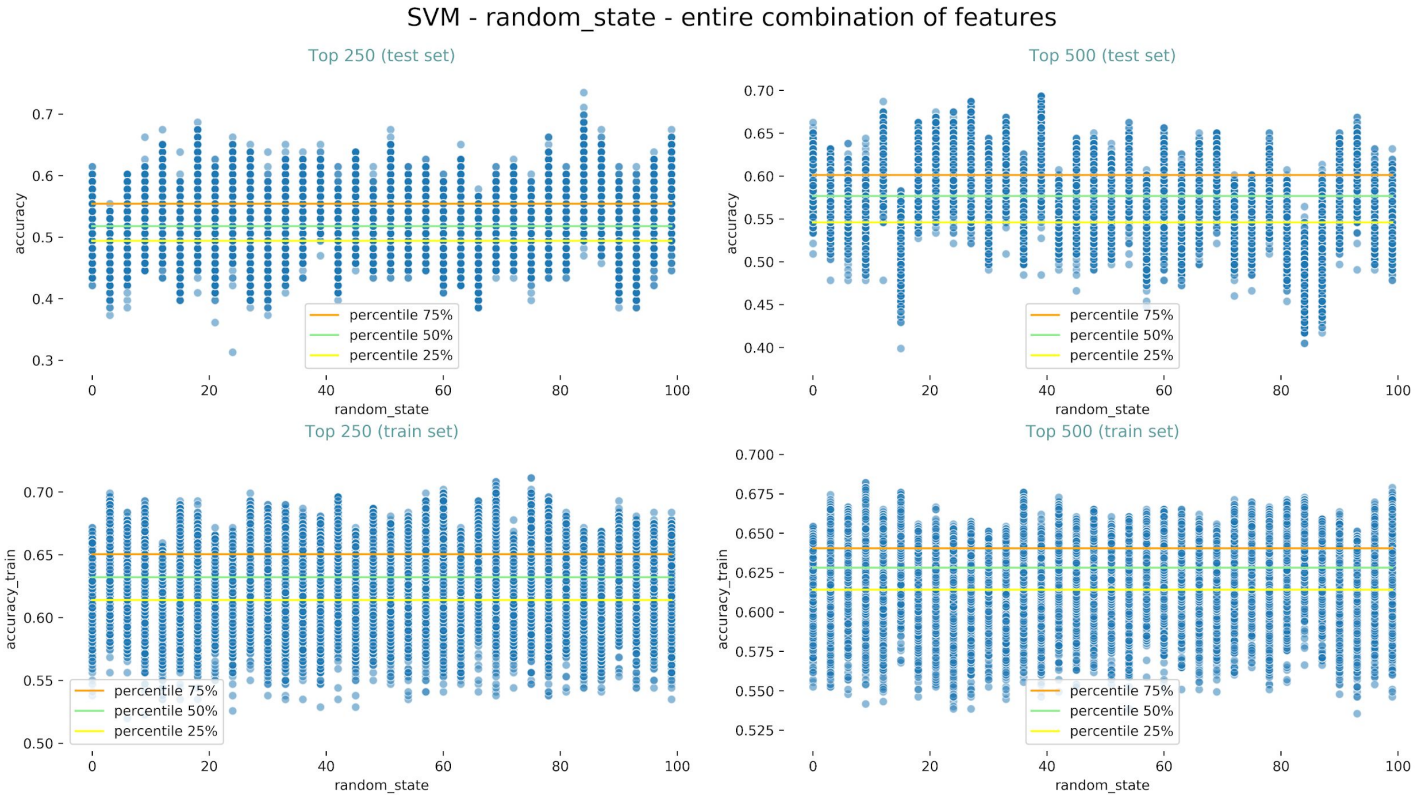
2 PREDICTION MODELS

I D M R D C METHODOLOGY

SUPPORT VECTOR MACHINE

K-NEAREST NEIGHBORS

DECISION TREE



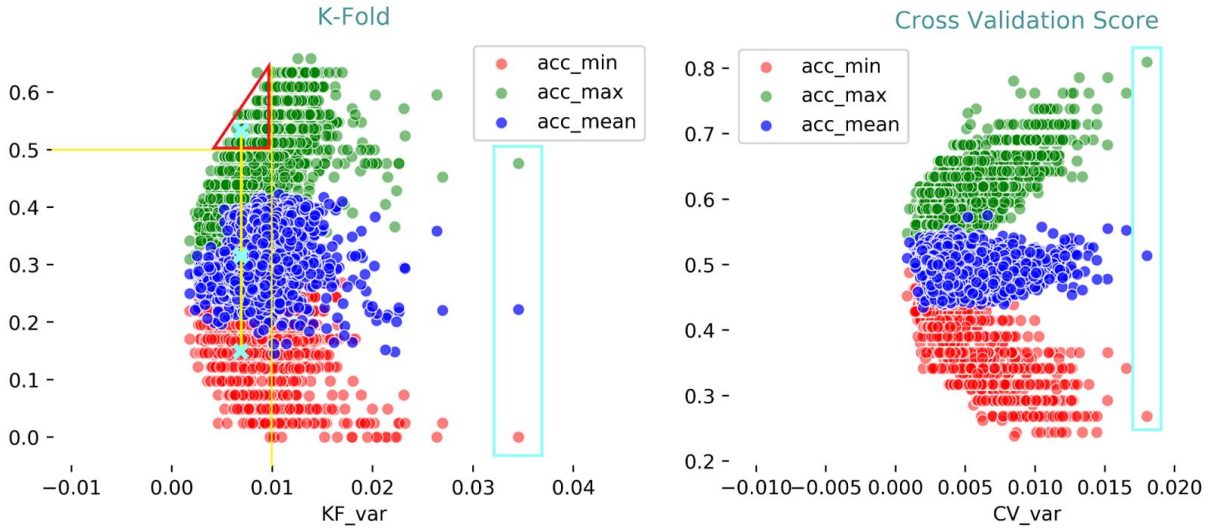
3 PREDICTION MODELS

I D R D C METHODOLOGY

SUPPORT VECTOR MACHINE

K-NEAREST NEIGHBORS

DECISION TREE



- In order to find the best combination of features for the model we must get the group with the highest accuracy and lowest variance for K-Fold and Cross Validation and see if we get a match
- For each single combination of features, point in the chart in a color group, there are mirrors in the other 2 groups, due to these points sharing the same variance

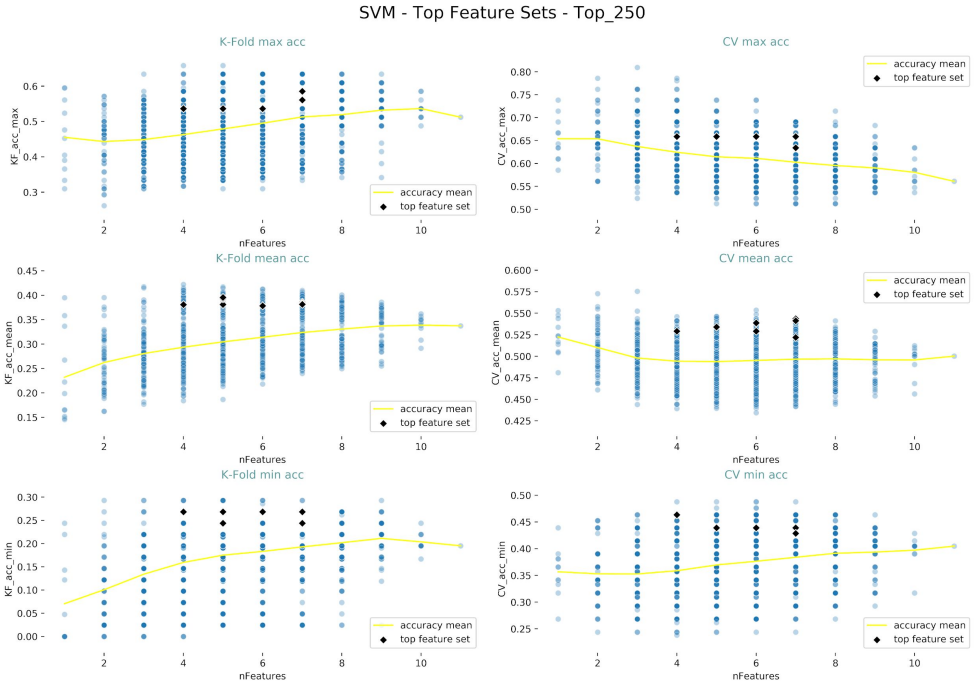
4 PREDICTION MODELS

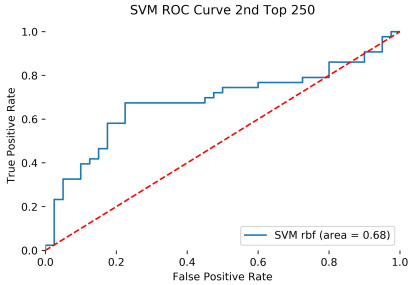
IDMRDC METHODOLOGY

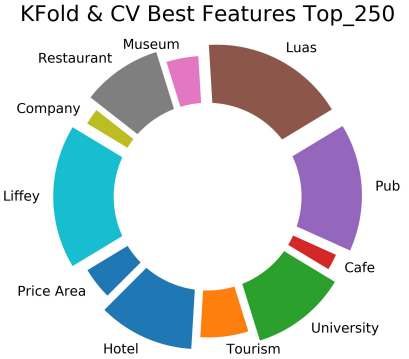
SUPPORT VECTOR MACHINE

K-NEAREST NEIGHBORS

DECISION TREE







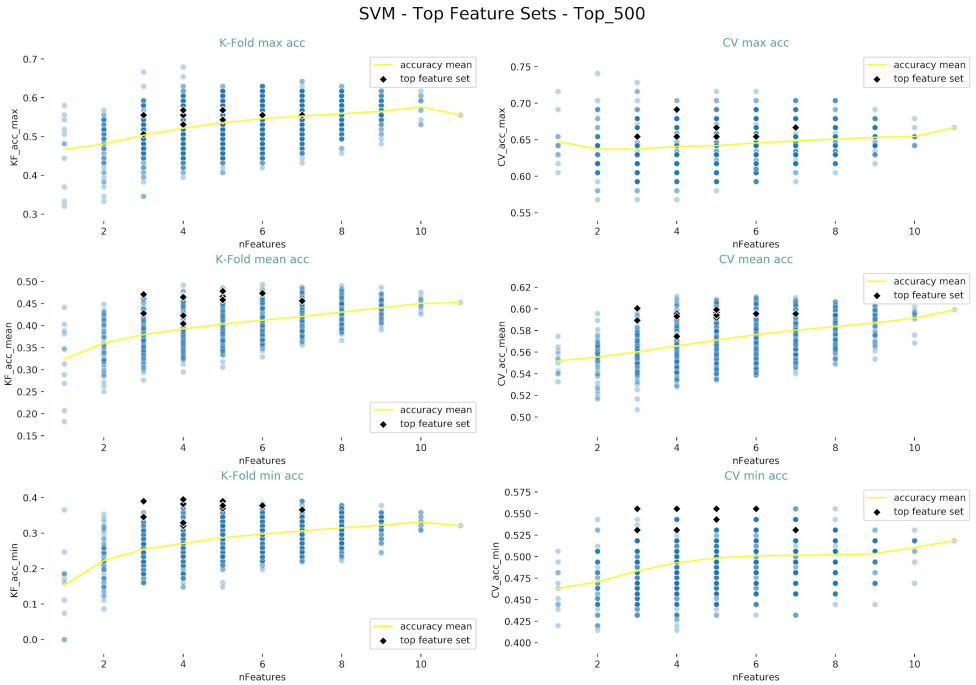
5 PREDICTION MODELS

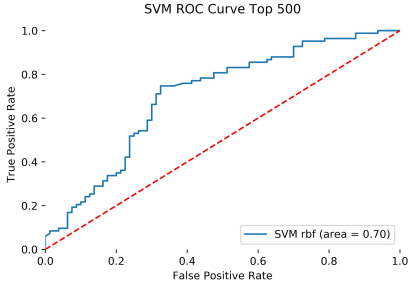
I D R D C METHODOLOGY

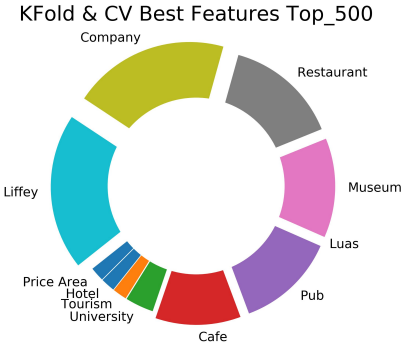
SUPPORT VECTOR MACHINE

K-NEAREST NEIGHBORS

DECISION TREE



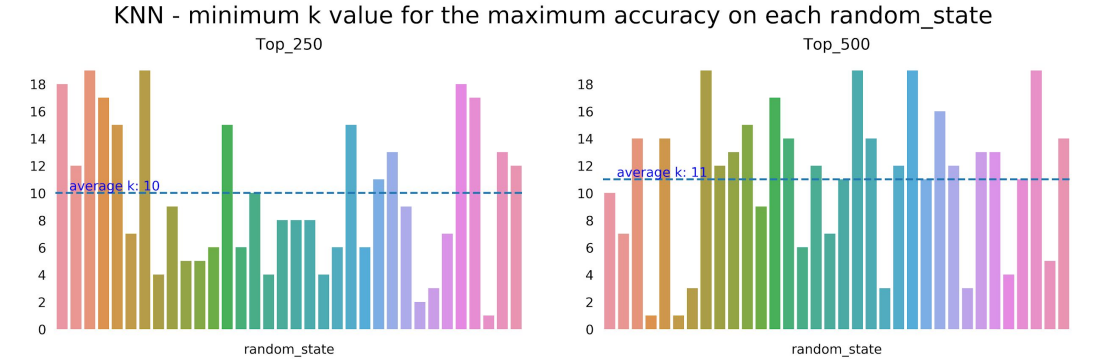


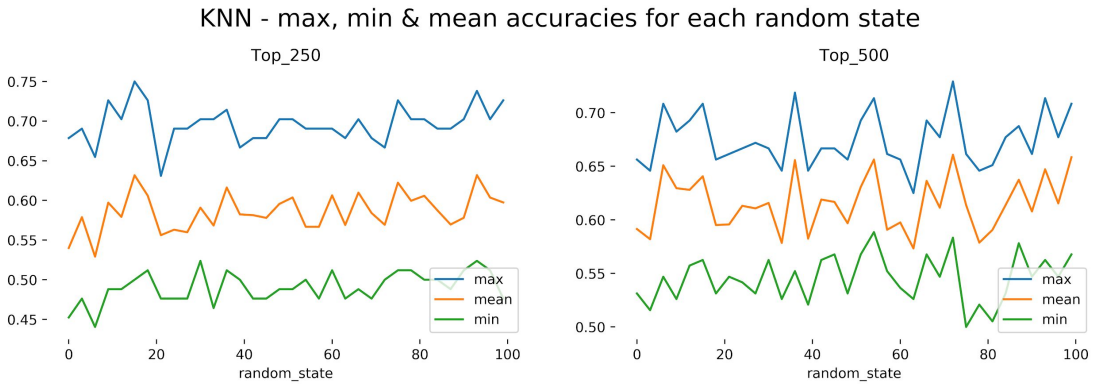


1 PREDICTION MODELS

I D M R D C METHODOLOGY

- SUPPORT VECTOR MACHINE
- K-NEAREST NEIGHBORS
- DECISION TREE





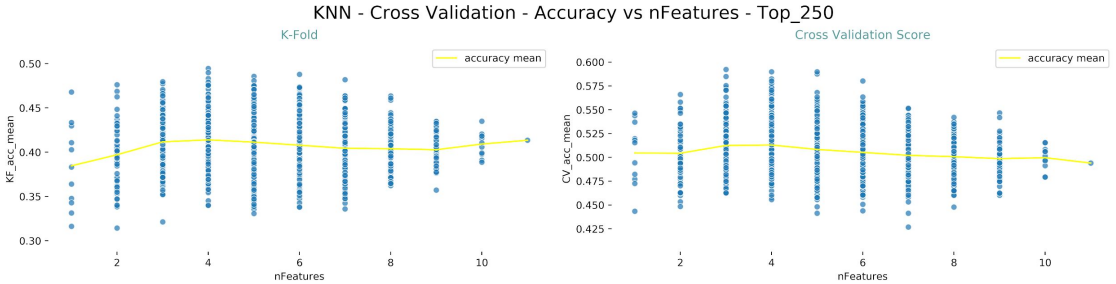
2 PREDICTION MODELS

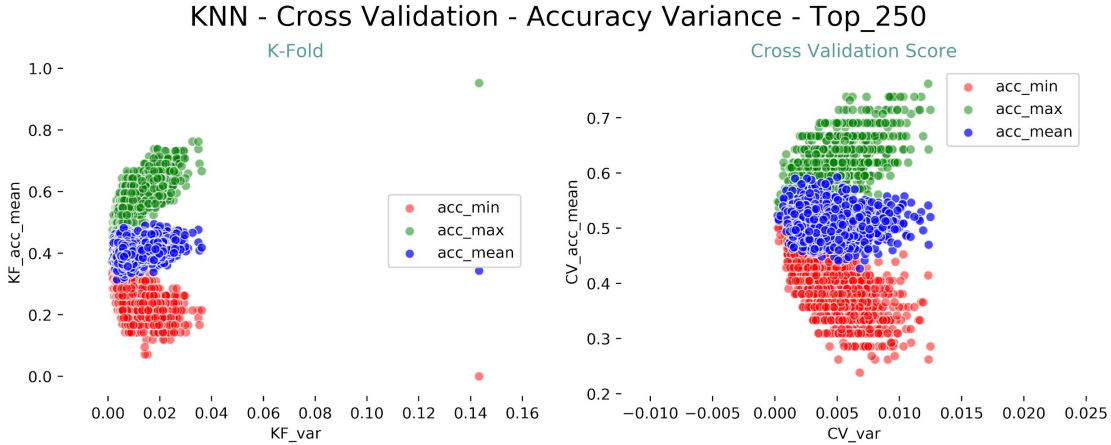
I D M R D C METHODOLOGY

SUPPORT VECTOR MACHINE

K-NEAREST NEIGHBORS

DECISION TREE

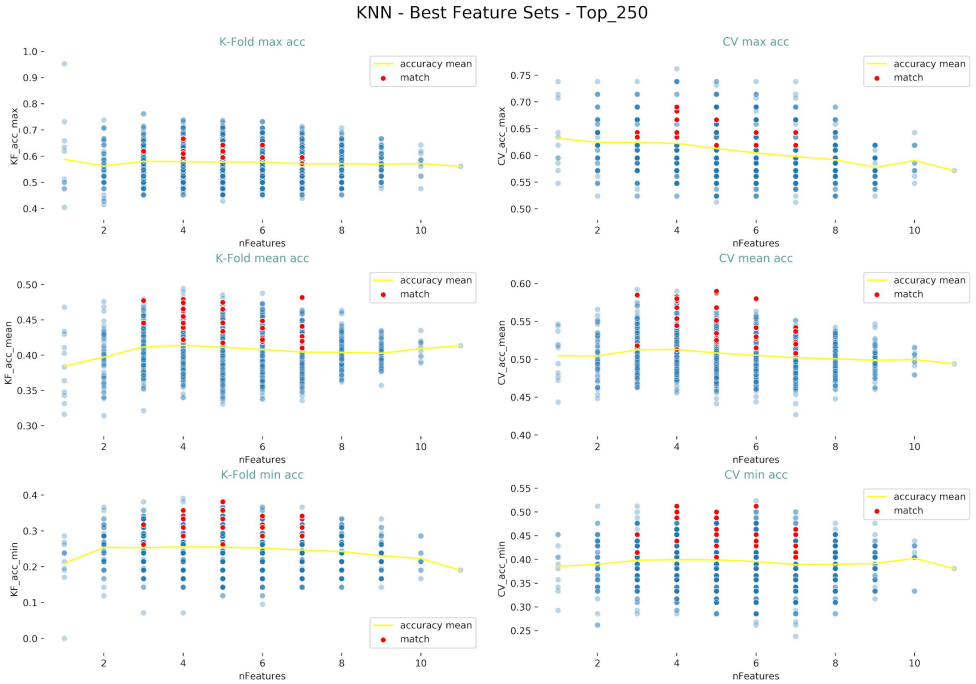


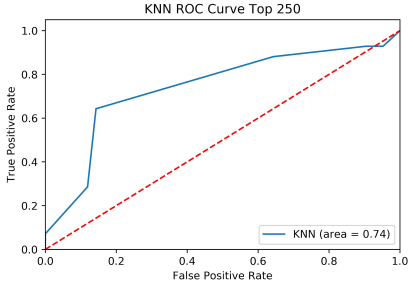


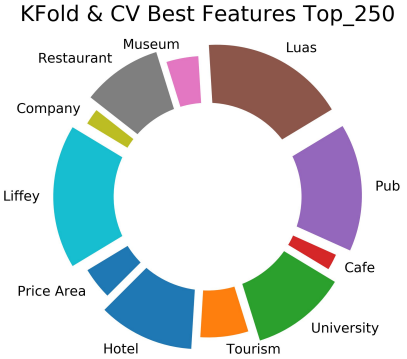
3 PREDICTION MODELS

I D M R D C METHODOLOGY

- SUPPORT VECTOR MACHINE
- K-NEAREST NEIGHBORS
- DECISION TREE







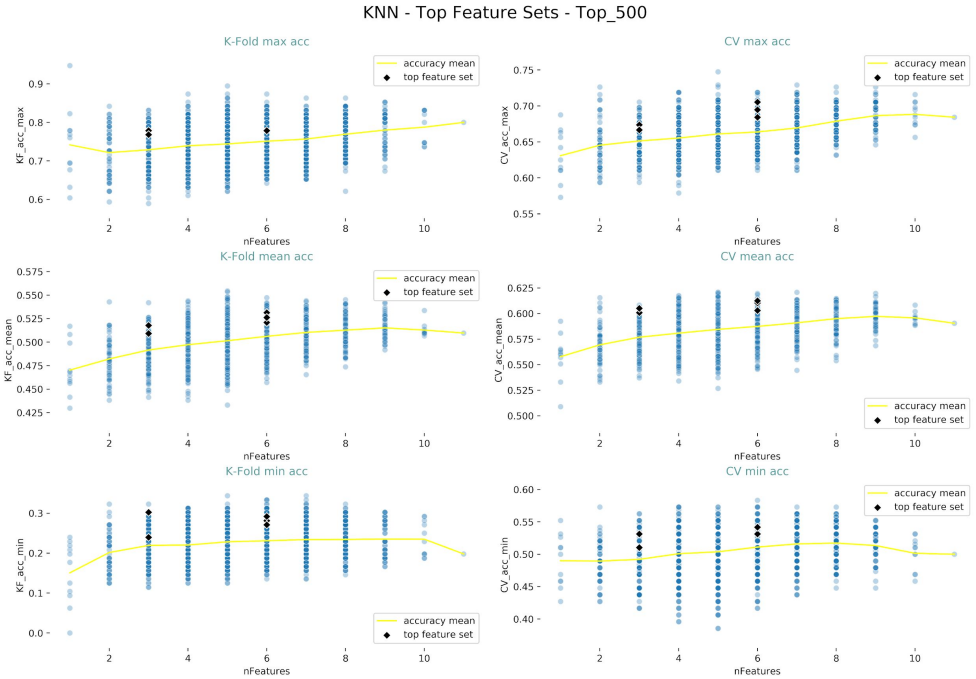
4 PREDICTION MODELS

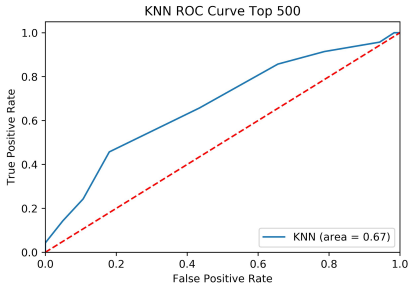
I D M R D C METHODOLOGY

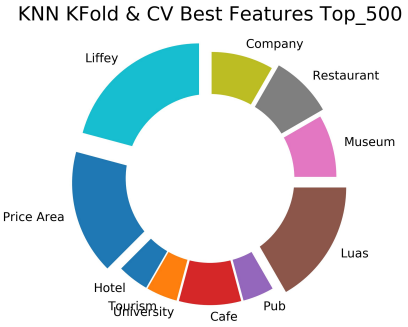
SUPPORT VECTOR MACHINE

K-NEAREST NEIGHBORS

DECISION TREE







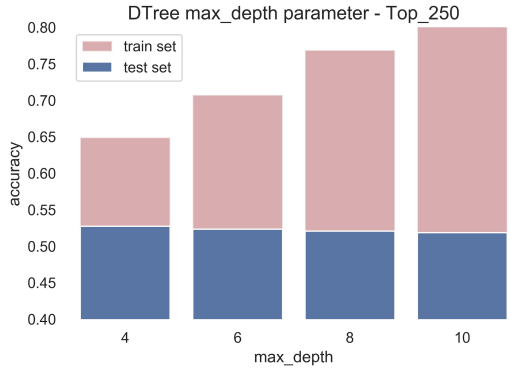
1 PREDICTION MODELS

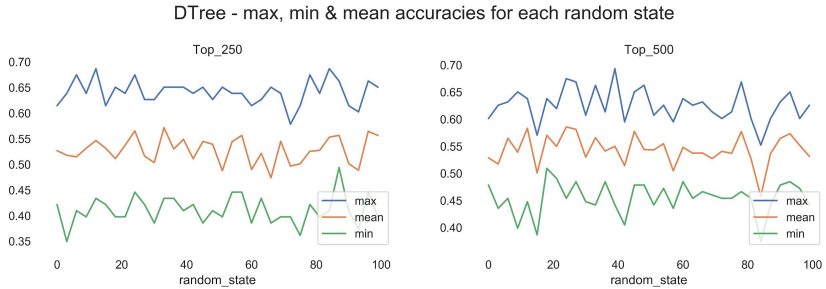
I D M R D C METHODOLOGY

SUPPORT VECTOR MACHINE

K-NEAREST NEIGHBORS

DECISION TREE





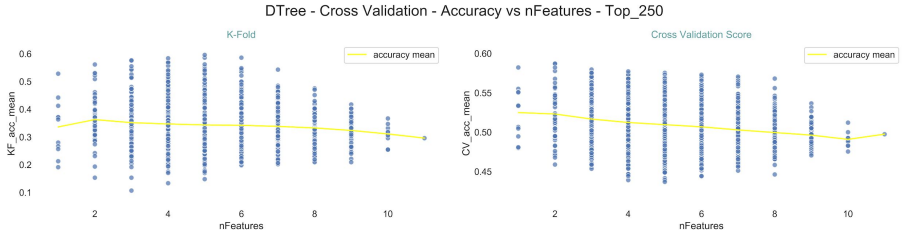
2 PREDICTION MODELS

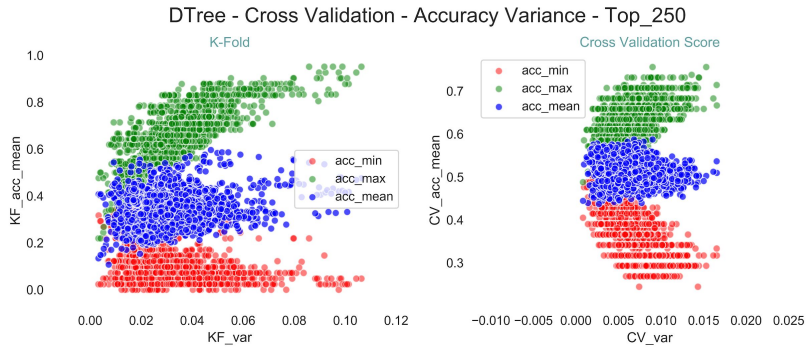
I D M R D C METHODOLOGY

SUPPORT VECTOR MACHINE

K-NEAREST NEIGHBORS

DECISION TREE





3 METHODOLOGY

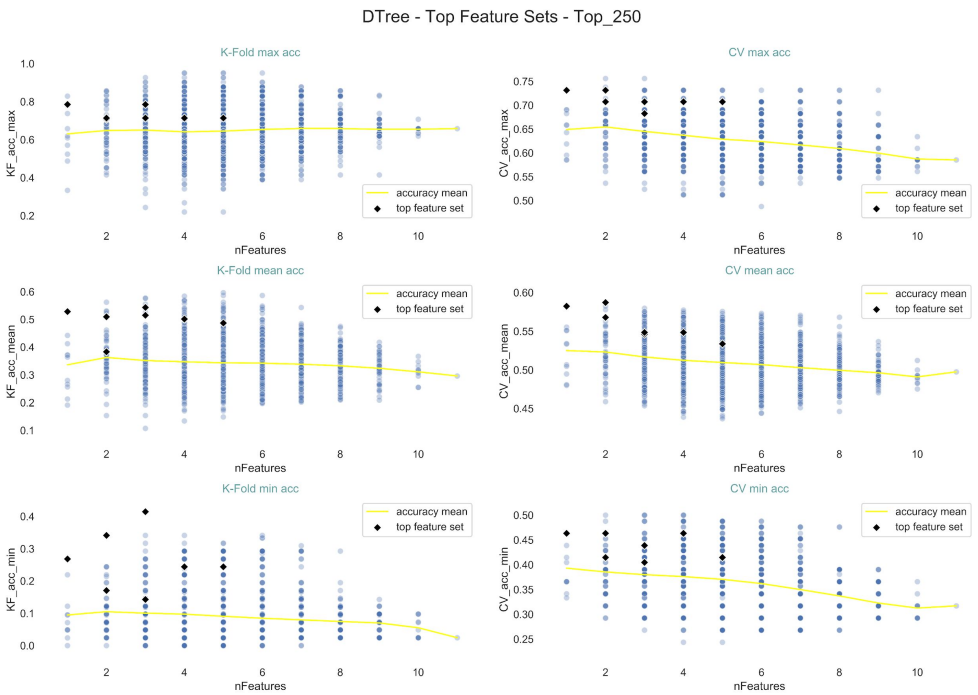
I D M R D C

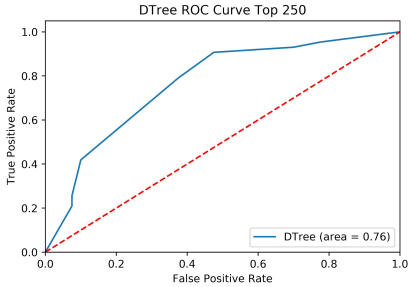
METHODOLOGY

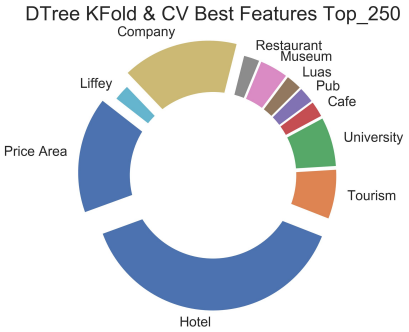
SUPPORT VECTOR MACHINE

K-NEAREST NEIGHBORS

DECISION TREE







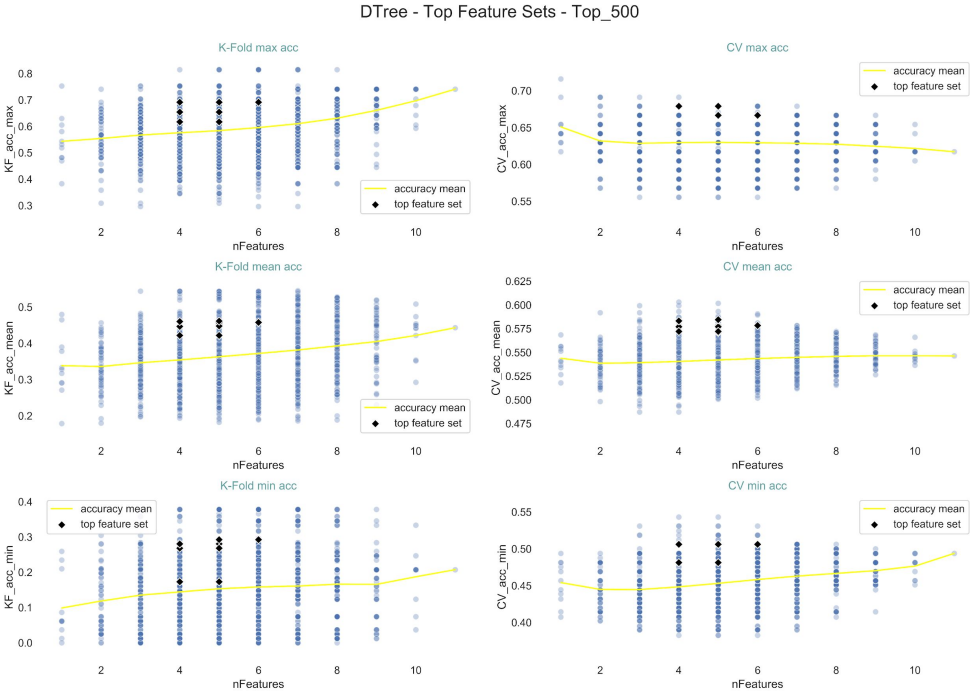
4 PREDICTION MODELS

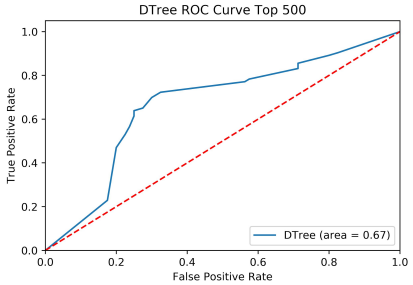
IDMRDC METHODOLOGY

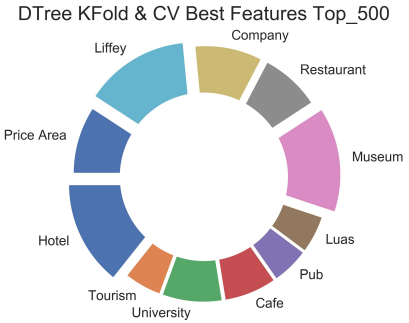
SUPPORT VECTOR MACHINE

K-NEAREST NEIGHBORS

DECISION TREE

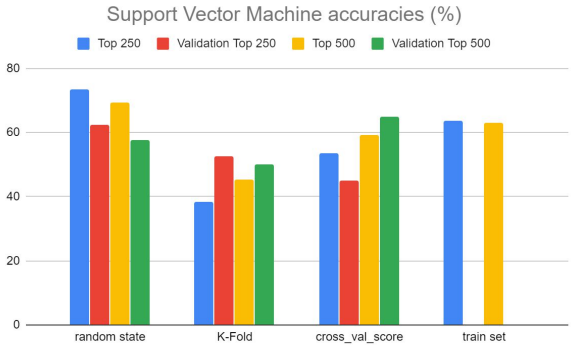


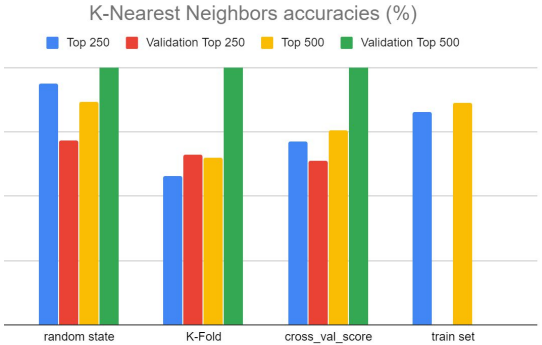


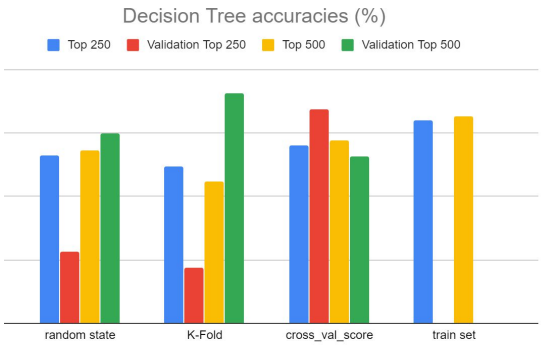


RESULTS

I D M R D C RESULTS







- Validation set above 80% on KNN
- Random_state approach average of 70% SVM & KNN
- Train set above 60%
- Top_250 above 70% on SVM % KNN

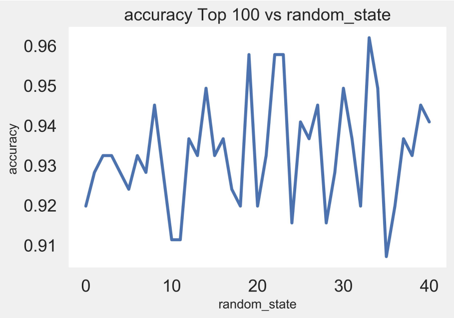
DISCUSSION

DISCUSSION

High acuracy on imbalanced target ⟹ Accuracy Paradox

cumulative ranking share

Top 100	5.3%
Top 250	18.9%
Top 500	45.4%
Top 750	81.1




```
yhat
```

```
array([0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
       0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
       0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
       0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
       0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
       0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
       0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
       0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
       0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
       0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
       0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
       0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
       0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
       0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.])
```

CONCLUSION

- Dublin has a nuclear distribution around the city center which has increased the difficulty in using classifiers as a means of demonstrating the capability of the model
- Centralised infrastructure of the city is having an important influence on how the algorithms behave
- As a follow on project, I believe it would be interesting to investigate this approach in a city with a different social infrastructure, for example, a city built within the last two centuries
- Even though, repeating the calculations for the Top 250 and Top 500 models increased the overall time spent, I was curious to see the final results for both, and which model would potentially be a better indicator for future projects

RESTAURANT LOCATION RESEARCH IN DUBLIN

IBM DATA SCIENCE CAPSTONE PROJECT PRESENTATION

JOSE HERNANDEZ

Links

Full Report

GitHub repository with the code
https://github.com/jotagar/IBM_Data_Science_Capstone

For a better user experience, I recommend using Jupyter nbviewer to access to the repository:
Code using nbviewer jupyter

Dublin
June 2020
jh.hdez@gmail.com